Leon Dappen
Jody C. Isernhagen
Planning and Changing
Vol. 36, No. 3&4, 2005, pp. 147–156

NEBRASKA STARS: ASSESSMENT FOR LEARNING

Nebraska's School-based, Teacher-led Assessment and Reporting System (STARS) is identified by the Partnership for the 21st Century Skills (2005) as "…the nation's most innovative assessment system" (p. 13). STARS is being watched closely by national audiences, but most importantly, it is described by a Nebraska school leader as "one of the best things we've done in my 25 years in education."

When confronted with No Child Left Behind (NCLB) and Average Yearly Progress (AYP) requirements, every state but Nebraska decided to use norm-referenced or state developed high-stakes measures. In a search for evidence of the positive effects of high-stakes tests on student achievement, Stiggins (2004) found only one study with small gains. He went on to describe a number of research studies that identified unintended but negative outcomes from high-stakes tests. One of the arguments against high-stakes assessment systems is that they focus on easily measured material leaving out less easily tested but possibly more important skills. They may also focus on lower level thinking and discourage creative activities. Recent studies have shown that while testing may result in score gains these gains are rarely lasting and are confined to the limited material being tested (Madaus, 1988; Haney, 2000). Standardized achievement tests measure students' innate abilities or background experiences and do not measure how well teachers teach or students learn. Gough (2000) reported that standardized tests tend to result in narrowing the curriculum, negatively affecting higher order thinking skills. Popham (1999) urged that standardized achievement tests should not be used as single measures of educational quality as they are not designed to match or measure state standards.

Madaus (1988) suggested that teacher-designed assessments and a focus on important student outcomes as identified by teachers may overcome the aforementioned assessment concerns. This system would give teachers opportunities to think through objectives and identify explicit types of evidence that would demonstrate that the objectives had been met. Jones and Ongtooguk (2002) suggested that assessments be multifaceted and based on student performance and teacher judgments. While it would be too simplistic to conclude that teacher designed assessment is a magic bullet for student achievement, it is clear that achievement can be raised only by changes that are put into effect by teachers and pupils (Black & William, 1998). As described by Neill, Guisbond, Schaeffer, Madden, and Legeros (2004), any new accountability system should include a local accountability system that provides teachers with high quality assessments that encompass a variety of ways to demonstrate knowledge and that fit with how children learn.

Why Nebraska STARS?

The philosophy of Nebraska STARS is based on the premise that the purpose of assessment is to drive curriculum and instruction to pro-

duce student academic achievement gain (Bandelos, 2004). STARS is a locally driven system designed from the classroom up which recognizes that improved student achievement will best occur with the focus on the interaction of teachers and students (Stiggins, 2004). Teacher designed standards, instruction, and assessments become part of a continuous improvement cycle. Based on this belief, Nebraska developed STARS to keep teaching and learning at the center of the educational process, promoting high-impact, not high-stakes, assessment (Gallager, 2004a).

What is Nebraska STARS?

Nebraska STARS requires each district either to adopt state standards or develop local standards that are at least equal to the state standards. Each district then develops a plan for assessing its standards. The plan is based primarily on locally developed criterion-referenced tests (CRTs), which are, therefore, unique to that district. STARS is completed at fourth, eighth, and eleventh grades. Since districts must address the quality assessment criteria (to be discussed later), which include that students have had the opportunity to learn the content prior to assessment, the timing of assessments is up to each district. Districts are also required to administer a standardized norm-referenced test (NRT) of their choosing (e.g., Terra Nova, Stanford Achievement Test) which provides an external common "touch point," and parts of which may also be used to assess some standards. The timing of the NRT administration is up to each district—once again addressing the issue of "opportunity to learn" for any standards being assessed by the NRT.

A unique aspect of STARS is a criterion-referenced statewide writing assessment based on the six-trait (see below) writing approach. With previous involvement by a number of local school districts in this writing model and the natural link between this criterion-referenced approach and the emerging philosophy of Nebraska STARS, a requirement for a statewide writing assessment was included in the legislation establishing Nebraska's assessment system. The Nebraska Department of Education coordinates with area Educational Service Units and local districts to implement the assessment at the same time across the state. A panel of teachers develops, refines, and pilots the prompts to be used at the fourth, eighth, and eleventh grades. Trained teachers come together at one site using rubrics developed for each grade level to holistically score the writing assessments. The scoring process and examination of results is carried out by the Buros Center for Testing, based in Lincoln, Nebraska. A sample of student papers from each grade is sent out of state for scoring by an independent testing company contracted for this purpose.

Description of District Assessment Portfolio

Since standards may be locally developed and the tests used to measure them are a mix of locally developed criterion-referenced measures, sections of district specific norm-referenced tests, and the statewide

writing assessment, there are few common measures to all districts. It must be remembered that STARS is designed to support instruction in local classrooms, not to facilitate ranking of schools. This strong reliance on district developed criterion-referenced measures challenges traditional validity and reliability views. Therefore, the primary measure of credibility for assessments is a District Assessment Portfolio that is submitted annually to the Nebraska Department of Education. The Portfolio includes school district ratings on six Quality Criteria that were identified by the Buros Center for Testing (Plake & Impara, 2000), the technical advisors to the STARS program. The six Quality Criteria require that: (a) assessments reflect state or local standards, (b) students have an opportunity to learn the content, (c) assessments are free from bias or offensive language or situations, (d) the level is appropriate for the students, (e) there is consistency in scoring, and (f) the mastery levels are appropriate.

Purpose of Study

The purpose of this study was to (a) examine criterion-referenced and norm-referenced student achievement data and District Assessment Portfolio ratings from STARS, and (b) interview school staff stakeholders regarding strengths, challenges, and recommendations for the STARS program.

Methodology

This was a mixed methods study with both quantitative and qualitative data.

Quantitative Methods

Achievement and portfolio data collection and description. The Nebraska Department of Education releases results each fall on the department website (Nebraska Department of Education, 2005a) that have been sent in from each district. The data include the district results on the statewide writing assessment, reading and math assessments, and the District Assessment Portfolio results. Local district and individual school data shared include: the percent of students meeting proficiency for criterion-referenced assessments (i.e., locally developed tests as well as the statewide writing assessment); the average percent of students in the top two quartiles on the district chosen norm-referenced test; and District Assessment Portfolio ratings.

Achievement data sample. Data were included for Class 3, 4, and 5 school districts. Class 3 school districts (209 school districts) are represented by any school district with territory having a population of more than 1000 but less than 150,000 inhabitants. Class 4 school districts (Lincoln only) have a population of 100,000 or more with a city of the primary class (between 100,000 and 200,000 inhabitants). Class 5 school districts

(Omaha only) have a population of 200,000 or more inhabitants with a city of the metropolitan class (over 300,000 inhabitants) within the territory (Nebraska Department of Education, 2004/2005). The districts in this study represent just over 94% of the public school students in Nebraska. The district data for this study were included on the state website, and cooperation for use of the data was facilitated by the Nebraska Department of Education.

District Assessment Portfolio ratings. District Assessment Portfolio ratings reflect each school staff's work in developing assessments that meet the six Quality Criteria. Portfolios are rated by an independent measurement expert specifically trained in the rubrics of each of the six Quality Criteria. The Buros Institute for Assessment Consultation and Outreach arranges for a panel of external reviewers comprising professionals with an earned doctorate in educational measurement. The rubric-based ratings on each criterion provide the basis for an overall rating. The overall rating scale ranges from "1," unacceptable, to "5," exemplary (Plake & Impara, 2000).

Qualitative Methods

Interview sample. From a survey sent statewide to 1,722 school stakeholders (superintendents; elementary and secondary principals; fourth, eighth, and eleventh grade math and language arts teachers; and Educational Service Unit staff developers), 25 districts were identified for follow up individual interviews. Districts were chosen based on geography, percent of free and reduced lunch students, and class size to represent the state as a whole. Permission was obtained from superintendents, and individual interviews were held with all staff available in the aforementioned categories. In total, 169 interviews were completed.

Interview questions and process. The statewide survey included demographic information, and questions related to strengths, challenges, and recommendations that individuals identified with respect to the STARS process. Interviews lasted from 15 minutes to an hour and probed interviewees further on the survey questions. The STARS Comprehensive Evaluation staff and three retired Nebraska administrators who had received training carried out the interviews using a common format. All interviews were taped and transcribed. The Comprehensive Evaluation staff then analyzed the interview transcription results for common themes.

Results

Quantitative Data

STARS reading achievement. Pre/post achievement data is available for various areas, grade levels, and years. The average percent of students reported by districts as mastering locally defined criterion-referenced reading tests from 2001 to 2003 increased 5.2% at fourth grade, .8% at

eighth grade, and 1.0% at eleventh grade. The average percent of students in the top two quartiles on the norm-referenced reading test used by districts increased 2.6% at fourth grade, decreased .6% at eighth grade, and increased 1.2% at eleventh grade.

STARS math achievement. Criterion-referenced math tests from 2002 to 2004 increased 6.6% at fourth grade, 6.5% at eighth grade, and 5.5% at eleventh grade. Norm-referenced math scores increased 3.3% at fourth grade, decreased .7% at eighth grade, and increased .1% at eleventh grade.

STARS writing achievement. The statewide writing assessment increased 4.41% from 2002 to 2004 at fourth grade; eighth grade increased 6.18% from 2003 to 2004. Eleventh grade was implemented statewide in 2004; post data are not yet available.

District Assessment Portfolio ratings. The average District Assessment Portfolio rating on the "1" to "5" Likert scale in reading across grades four, eight, and eleven, increased from 3.5 in 2001 to 4.35 in 2003. The average District Assessment Portfolio rating across grades in math increased from 3.97 in 2002 to 4.74 in 2004. The statewide writing test does not require a separate District Assessment Portfolio. In looking at District Assessment Portfolio ratings from year to year across reading and math, the average rating increased from 3.5 in 2001, to 3.97 in 2002, to 4.35 in 2003, to 4.74 in 2004.

Qualitative Data

Interviewee described strengths of STARS. Interviews of stakeholders (school board presidents, superintendents, elementary and secondary principals, fourth, eighth, and eleventh grade reading and math teachers, and educational service unit staff developers) have provided generally strong, however, recognizably mixed support for the program. Interviews revealed general growth in teachers' assessment literacy and use of student data in refining instruction and curriculum planning. Educators reported that, over the years, specific concerns have been identified and changes and improvements have been made (Gallagher, 2004b; Isernhagen, 2005). A superintendent commented:

> With the implementation of the STARS process, educators are incorporating data with instruction. Communication about student learning is at an all-time high. Collaboration among teachers has increased. Assessments have greatly improved and their importance has increased. The time and effort is well worth the outcome.

A teacher commented, "Assessments are very useful for our local school improvement activities."

The STARS process is described as matching well with research regarding best practice in school improvement and staff development, as described by Supovitz and Christman (2005), and in providing an impetus

for professional learning teams in schools. Interviews with teachers revealed strong support for the process. One teacher commented, "When we sat down to look at our math improvement plan, we had the reports from various testing to see where we scored low or high and we talked about what we might do to bring about change." The process embodies differentiated leadership with administrators working with teacher leaders and teams of teachers on direct teaching and learning activities. The focus is on data-based instructional improvement rather than ratings and rankings.

Interviewee described challenges related to STARS. The strongest and most often expressed concern regarding STARS is the time to support activities. Even those very supportive of the program express concern regarding the time involved. A teacher stated, "The assessments have good points. The downside is the time it takes from teaching and preparation." Another said, "It all boils down to time." Teachers and principals who have been directly involved in developing and administering assessments comment on the high value in improving learning, however, sometimes feel overburdened. Staff developers involved with STARS training reported that they have been able to reduce the training time necessary for a "cycle" of an academic area by one-half or more since the program began. An area agency trainer stated, "I can see the difference as the same teachers come back to work on new areas, they work much quicker." Since many common elements are involved, experienced staff will be able to complete "cycles" in less time in the future. While time continues to be of great concern, increases in District Portfolio ratings reflect increased assessment competence, and comments from trainers indicate increased efficiency in training as the program has progressed.

A related concern is the number of staff to train. In small districts, the training to enable implementation of the program results in a high percentage of the staff being trained. In large districts, where there are many more staff to train, trainer of trainer models and other efforts have been employed. The time involved in these training issues also translates into significant financial and strategic concerns for districts (i.e., when to train with the least negative effect on student contact time).

Interviewee described issues with NCLB. Nebraska is using STARS to comply with No Child Left Behind requirements that focus on norm-referenced test gains. Because of STARS' unique nature and different philosophy, considerable challenges have resulted. While Nebraska's Commissioner of Education and Department of Education have worked long and hard, the compromises made to gain federal approval for STARS have created new challenges. Discussion with the U.S. Department of Education to gain approval of STARS for NCLB purposes has resulted in expansion of the grade levels to be tested. Furthermore, many districts had recognized the value and importance of criterion-referenced tests prior to NCLB and, with STARS, already had comprehensive criterion-referenced assessment programs in development or in place. The inflexibility of

NCLB testing requirements has resulted in much time focused on rewriting and refining criterion-referenced measures to reflect higher "technical characteristics," to be administered at more grade levels, and to meet other requirements for purposes of NCLB. This has greatly added to the time and cost of the process with little added to teaching and learning. Though most agree that the STARS model is best for teaching and learning, some Nebraskans feel "beat down" by the federal requirements and have discussed going to a single state test. A superintendent, concerned about the challenges of STARS resulting in educators wanting a single test, commented "we continue to encourage leadership of the state to not fall victim, and so far so good." Most educators feel that those states using state tests have similar concerns regarding NCLB, and they have little confidence that a decision to implement a state test would be received any better or cost any less than STARS. As one administrator commented, "AYP is very cumbersome but the STARS process seems to be a better system than a one test approach." Another commented, "I hope Nebraska never moves toward a standardized test." A further concern expressed was that district or state designed assessments would, as many suggest with norm-referenced tests, result in narrowing the curriculum to the test. In spite of these concerns, Hillocks (2002), in examining state developed writing tests in Illinois, Texas, New York, Kentucky, and Oregon, reported that while teachers believed their systems may narrow the curriculum, the programs supported a desirable writing program and improved student writing.

Interviewee described measurement concerns. A final challenge is a lack of understanding of criterion-referenced measures and how they may be used in a statewide program. Likely because of the many years of employing statistics, the measurement community has been reasonably effective in creating some degree of understanding regarding statistics and their use in norm-referenced measures. There is a significant need to focus on criterion-referenced measures and statistics that may be used in statewide programs such as this. Traditional statistical applications are insufficient for these efforts (Isernhagen & Dappen, 2005). An initial effort to explore this issue has been discussed at the first annual conference on Leadership in Classroom Assessment that was held in Omaha, Nebraska, in September of 2005. Information regarding future conferences can be obtained from the Nebraska Department of Education (2005b).

Discussion and Conclusions

Criterion-referenced measures and the statewide writing test are showing decent growth. The stronger growth in math and writing may be based on the extra year of training and experience that schools have had with the process; future longitudinal results will provide more information regarding this point. Norm-referenced measures have generally increased, though not as much. This has been a positive finding since there was concern from educators as to whether the attention focused on criterion-refer-

enced measures might result in a decline in traditional norm-referenced measures. The independent professional ratings of Assessment Portfolios reveal strong, consistent growth in district staff abilities to create assessments that meet the Quality Criteria identified as the backbone of the STARS system and essential for the program's credibility.

Real school improvement with student academic achievement as the goal is not a short-term process. Those looking for striking success measures in the short run will be disappointed. Nebraska is in the fifth year of STARS implementation and comments from very positive supporters would indicate that we are still several years from full implementation of the program. As one administrator commented, "I believe we will get there, but it will take a few more years." It is clear that changing the paradigm—to focus on how data from criterion-referenced measures can impact curriculum and instruction to achieve academic student gain through continuous school improvement cycles based on professional learning communities—is a tall order and requires significant commitment, resources, and time. But it is happening.

While schools relentlessly pursue academic gain for all students, we must recognize some realities. Children who come to school with one-third of the vocabulary of others, who have vision or hearing problems resulting from little or no basic health care, who move two or three times a year, or who enter school speaking English for the first time in their lives require a lot more than that they be tested and expected to be at grade level in a certain time frame (Rothstein, 2004). To lay this challenge on schools by edict with no money and no plan makes educators extremely skeptical. No assessment system is going to ensure achievement of that goal; yet, schools are being measured by that expectation and pronounced "failing" if they are not on target to achieve it. Educators do not support excuses, but, as stated by one superintendent, "because we aren't perfect doesn't mean we are failing." We (Americans) owe students a comprehensive, coordinated, and funded plan involving all appropriate agencies to address the myriad of issues involved.

Effective and efficient methods for implementation of the STARS model have been developed, are in use, and the data are positive. Ongoing evaluation has revealed concerns that have been or are being addressed. Interviews of stakeholders revealed increasing assessment literacy and application. The STARS model fits into best practice for professional development, assessment, and approaches for student academic gain. The "front end" hard work is paying off, and we can see long-term gain as realistic. As one school leader summarized, "The overwhelming topics of discussion in Nebraska schools revolve around teaching and learning." The concern most often expressed by those involved with the Nebraska STARS program is for recognition from the federal government of STARS being a credible alternative and deserving of flexibility in implementation. If the federal government seeks new and effective alternatives in public education, it needs to give more flexibility to public schools, as well as to other types of schools, for promising alternatives to be fully explored.

References

Bandelos, D. L. (2004). Introduction to the special issue on Nebraska's alternative approach to statewide assessment. *Educational Measurement: Issues and Practice, 23*(8), 6–8.

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessments. *Phi Delta Kappan, 79*(3), 139–148.

Gallagher, C. (2004a). *Charting STARS: New conversations* (Year Three Report to Nebraska Department of Education). Lincoln: University of Nebraska.

Gallagher, C. (2004b). Turning accountability tables: Ten progressive lessons from one "backward" state. *Phi Delta Kappan, 86*(4), 352–360.

Gough, P. (2000). Moving toward the danger. *Phi Delta Kappan, 81*(9), 562–568.

Haney, W. (2000). The Texas miracle in education. *Education and Policy Analysis Archives*. Retrieved March 9, 2005, from: http://epaa.asu.edu/epaav8n1.html

Hillocks, G. (2002). *The testing trap: How state writing tests control learning.* New York: Teachers College Press.

Isernhagen, J. (2005). *Charting STARS: Voices from the field* (Year Four Report to Nebraska Department of Education). Lincoln: University of Nebraska.

Isernhagen, J., & Dappen, L. (2005, September). *Nebraska STARS: Evaluation of a statewide system.* Paper presented at the first annual conference of Leadership in Classroom Assessment, Omaha, NE.

Jones, K., & Ongtooguk, P. (2002). Equity for Alaskan natives. Can high-stakes testing bridge the chasm between ideals and realities? *Phi Delta Kappan, 83*, 499–513.

Madaus, G. F. (1988). The influence of testing on curriculum. In L. N. Turner (Ed.), *Yearbook of National Society for the Study of Education: Vol. 87. Critical issues in curriculum* (Part 1, pp. 83–121). Chicago: University of Chicago Press.

Nebraska Department of Education (2005a). *Accountability/STARS*. Retrieved March 9, 2005, from the Nebraska Department of Education Web site: http://www.nde.state.ne.us

Nebraska Department of Education (2005b). *Meetings and Conferences.* Retrieved August 17, 2005, from the Nebraska Department of Education Web site: http://www.nde.state.ne.us

Nebraska Department of Education (2004/2005). *Nebraska Education Directory* (13th ed.). Lincoln, NE: Author.

Neill, M., Guisbond, L., Schaeffer, B., Madden, J., & Legeros, L. (2004). *Failing our children: How No Child Left Behind undermines quality and equity in education.* Cambridge, MA: The National Center for Fair and Open Testing.

Partnership for 21st Century Skills. (2005). *The Road to 21st century learning: A policymaker's guide to 21st century skills.* Retrieved May 4, 2005, from http://www.21stcenturyskills.org/route/

Plake, B. C., & Impara, J. C. (2000). *Technical quality criteria for Nebraska's district assessments.* Lincoln, NE: Buros Center for Testing.

Popham, W. J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership, 56*, 8–15.

Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the Black-White achievement gap.* Washington, DC: Economic Policy Institute.

Stiggins, R. (2004). New assessment beliefs for a new school mission. *Phi Delta Kappan, 86*(1), 22–27.

Supovitz, J. A., & Christman, J. B. (2005). Small learning communities that actually learn: Lessons for school leaders. *Phi Delta Kappan, 86*(9), 649–651.

Leon Dappen is an Assistant Professor in the Department of Educational Administration and Supervision at the University of Nebraska at Omaha.

Jody C. Isernhagen is an Associate Professor in the Department of Educational Administration and Supervision at the University of Nebraska at Lincoln.